UTFPR at SemEval-2021 Task 1: Complexity Prediction by Combining BERT Vectors and Classic Features

Gustavo Henrique Paetzold
Universidade Tecnológica Federal do Paraná - Campus Toledo
Toledo-PR, Brazil
ghpaetzold@utfpr.edu.br

Abstract

We describe the UTFPR systems submitted to the Lexical Complexity Prediction shared task of SemEval 2021. They perform complexity prediction by combining classic features, such as word frequency, n-gram frequency, word length, and number of senses, with BERT vectors. We test numerous feature combinations and machine learning models in our experiments and find that BERT vectors, even if not optimized for the task at hand, are a great complement to classic features. We also find that employing the principle of compositionality can potentially help in phrase complexity prediction. Our systems place 45th out of 55 for single words and 29th out of 38 for phrases.

1 Introduction

Accurately measuring the complexity of words can be useful in many ways. It facilitates the creation of text simplification technologies that could, for example, help in identifying and adapting challenging excerpts of literary pieces targeting specific groups, such as children (De Belder and Moens, 2010) and second language learners (Paetzold and Specia, 2016e), and make news articles and official documents more accessible to the general population (Paetzold and Specia, 2016a).

This task has received a considerable amount of attention in the past few years, especially due to the popularity of the Complex Word Identification (CWI) shared tasks of 2016 (Paetzold and Specia, 2016c) and 2018 (Yimam et al., 2018), where dozens of teams were challenged to judge the complexity of words in context. While the CWI 2016 task used a simple binary complex/not complex classification setup for English only, the CWI 2018 task explored both a binary classification and a regression setup and multiple languages.

The majority of the most successful systems submitted to these shared tasks combined ensemble methods, such as Random Forests (Ho, 1995) and AdaBoost (Freund and Schapire, 1997) with numerous linguistic features, including word frequencies, n-gram frequencies, word length, number of senses, number of syllables, psycholinguistic metrics, and word embeddings (Konkol, 2016; Malmasi et al., 2016; Paetzold and Specia, 2016d; Gooding and Kochmar, 2018; Hartmann and Dos Santos, 2018). However, because these tasks were held prior to the ascension of transformer-based masked language models, such as BERT (Devlin et al., 2019) and ROBERTA (Liu et al., 2019), we could not find any systems that exploited the power of the features produced by them.

In this paper, we describe the UTFPR systems for the Lexical Complexity Prediction shared task of SemEval 2021 (LCP 2021), which combine classic complexity prediction features with contextual word and phrase representations extracted from transformer-based models. In our experiments, we explore the efficacy of a number of different machine learning models, feature combinations, and corpora sources for our features. In what follows, we present the task being addressed (Section 2), our approach (Section 3), some preliminary experiments (Section 4), our final shared task results (Section 5), and our conclusions (Section 6).

2 Task Description

We address the LCP 2021 shared task (Shardlow et al., 2021), held at SemEval 2021. The shared task is split into two sub-tasks: predicting the in-context lexical complexity of single words and phrases for the English language. Participants could choose to submit systems to either or both sub-tasks.

The organizers provided training, trial and test sets for both sub-tasks. Each instance of these datasets is composed of an ID, a source identifier,

Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pages 617–622
Bangkok, Thailand (online), August 5–6, 2021.

a sentence, a target word or phrase within the sentence, and a complexity score calculated based on judgments made by 20 English speakers from the USA, UK and Australia. The source identifier describes from where the sentence came from, the possibilities being the Bible, biomedical documents and the Europarl corpus. The task's dataset is an extended version of the CompLex dataset (Shardlow et al., 2020).

The training, trial, and test sets for single words have 7662, 421, and 917 instances, respectively. The training, trial and test sets for phrases have 1517, 99, and 184 instances, respectively. Participants were allowed and encouraged to use any external resources they saw fit.

3 Approach

Our approach consists of using modern ensemble models to learn from a combination of commonly used complexity estimation features, such as word frequencies, word length, and number of senses, with contextual representations extracted from large pre-trained BERT-like models, which have been widely used to create state-of-the-art solutions to numerous tasks. While it has been observed that word frequencies (especially those extracted from spoken text) tend to drive the performance of effective complexity prediction systems (Paetzold and Specia, 2016c), we hypothesize that the wealth of knowledge present in transformer-based models such as BERT can help in extracting complementary contextual complexity clues.

3.1 Features

We explore a set of 779 total features in our approach. They are:

- **Frequency:** We use not only word/phrase frequency, but also n-gram frequencies as well. We consider a total of 9 configurations (i, j), where i represents the number of tokens to the left of the target word/phrase to be considered and j the number of tokens to the right. The configurations we consider are (0, 0), (0, 1), (1, 0), (1, 1), (0, 2), (2, 0), (1, 2), (2, 1), (2, 2).
- **Length:** We use the number of characters that compose the word/phrase. For phrases, instead of using its overall length, we use the average number of characters of all individual words. We motivate this decision in the experiments of Section 4.2.
- **Number of senses:** We use the word/phrase's number of senses catalogued in the WordNet database (Miller et al., 1990). In line with our setup for word length, for phrases, we use the average number of senses of all individual words.
- **BERT vector:** We use the numerical representation of 768 dimensions produced by the pre-trained BERT model (Devlin et al., 2019). For phrases and out-of-vocabulary words that were fragmented during tokenization, we average the representations produced for all fragments. More specifically, we used the bert-base-uncased model from the Hugging Face's transformers library (Wolf et al., 2020).

In the experiments of Section 4.3, we conduct an ablation study that reveals the performance impact of adding/removing some of these features from our models.

3.2 Models

We explore 5 different machine learning models in our experiments:

- Ridge Regression (Ridge) (Tikhonov, 1943)
- Support Vector Machines (SVM) (Boser et al., 1992)
- AdaBoost Regression (AdaBoost) (Freund and Schapire, 1997)
- Gradient Boosting (GBoost) (Friedman, 2001)
- Random Forests (Forests) (Ho, 1995)

The final configuration we chose to submit to LCP 2021 is described in Section 5. In the following section, we explain how we got to that configuration.

4 Preliminary Experiments

In this section, we describe the preliminary experiments we conducted in an effort to engineer our final systems for the LCP 2021 shared task. In these experiments, all machine learning models were trained and optimized on the training set and tested on the trial set provided by the organizers. All models were implemented using the Scikit-Learn library (Pedregosa et al., 2011) and optimized using grid search and 5-fold cross validation.

Split	Size	(0, 0)	(0, 1)	(1, 0)	(1, 1)	(0, 2)	(2, 0)	(1, 2)	(2, 1)	(2, 2)
Chi-M	1.5M	0.569	0.302	0.288	0.229	0.254	0.211	0.205	0.183	0.171
Chi-S	1.5M	0.547	0.297	0.287	0.225	0.246	0.215	0.200	0.183	0.170
Chi-MS	3M	0.578	0.316	0.312	0.245	0.261	0.233	0.217	0.200	0.184
Fam-M	2.9M	**0.609**	0.334	0.318	0.255	0.284	0.232	0.232	0.201	0.191
Fam-S	3.1M	0.578	0.338	0.305	0.253	0.277	0.226	0.225	0.202	0.187
Fam-MS	6M	0.607	0.346	**0.327**	0.263	**0.291**	0.241	**0.239**	0.211	0.198
Com-M	19.3M	0.591	0.333	0.323	0.258	0.280	0.241	0.233	0.210	0.196
Com-S	15.7M	0.578	0.351	0.319	0.268	0.279	0.234	0.229	0.211	0.189
Com-MS	35M	0.571	0.339	0.323	0.264	0.277	0.243	0.233	0.216	0.197
Movies	21M	0.592	0.335	0.327	0.262	0.281	0.244	0.236	0.213	0.198
Series	17M	0.576	**0.352**	0.321	**0.269**	0.281	0.238	0.233	0.214	0.193
All	38M	0.570	0.341	0.326	0.267	0.279	**0.247**	0.236	**0.219**	**0.201**

Table 1: Trial set Pearson correlations between complexity scores and frequencies for all SubIMDB splits and n-gram configurations on the single words sub-track.

4.1 Corpora Analysis

Arguably the most important features we use are frequencies. These must be calculated based on a language model trained on a specific corpus, so, as a first step in our engineering process, we decided to conduct an experiment to choose a corpus for the shared task in question.

As evidenced and discussed by Brysbaert et al. (2012) and Paetzold and Specia (2016b), frequencies extracted from spoken text corpora tend to correlate better with word complexity, so we decided to choose the SubIMDB corpora (Paetzold and Specia, 2016b) for our experiment. SubIMDB is a structured corpus extracted from 38,102 subtitles of children, family and comedy movies and series. We created 12 SubIMDB splits for this experiment: Children movies (Chi-M), children series (Chi-S), children movies and series (Chi-MS), family movies (Fam-M), family series (Fam-S), family movies and series (Fam-MS), comedy movies (Com-M), comedy series (Com-S), comedy movies and series (Com-MS), all movies (Movies), all series (Series), and the entire corpus (All). We calculate the Pearson correlation between the trial set complexity scores and n-gram frequencies for all n-gram configurations described in Section 3.1. To do so, we trained 5-gram language models over these splits using KenLM (Heafield, 2011).

The results illustrated in Table 1 are absolute correlation scores for the trial set of the single words sub-track (original values were negative, given that complexity inversely correlates with word frequency). We chose absolute scores to make the table more compact. It can be observed that the (0, 0) configuration (no context) yields the best correlations in every scenario. It can also be noted that, while the family movies split (Fam-M) is best for (0, 0), the remaining configurations tend to benefit from larger splits. Based on that observation, in the experiments that follow, we use family movies to calculate frequencies for single words/phrases and the whole SubIMDB corpus for the remaining n-grams.

4.2 Phrase Compositionality

The next step in our engineering process was to optimize the performance of our submission for the phrases sub-track. For that, we tested the hypothesis that the complexity of a phrase can be more reliably modelled if addressed as a product of the complexity of its words. To do so, we first calculated 3 features from our feature set using 4 different composition functions, then calculated the Pearson correlation between them and the reference complexity scores from the trial set.

The features calculated are: Phrase/word frequency, length, and number of senses. The composition functions are: None (addressing the phrase as a single word), averaging, maximum, and minimum. Frequencies were calculated using a 5-gram language model trained over the entire SubIMDB corpus.

The results in Table 2 show that, overall, employing the principle of compositionality in feature calculation for phrases increases the correlation between classic complexity features and human complexity scores. This is especially true for word senses, given that Wordnet has very few phrases

	None	Avg.	Max.	Min.
Frequency	−0.617	−0.641	**−0.650**	−0.547
Length	0.482	**0.482**	0.370	0.461
Senses	−0.125	**−0.460**	−0.430	−0.420

Table 2: Trial set Pearson correlations for different compositionality settings on the phrases sub-track.

catalogued.

In the subsequent experiments, we employ averaging as the compositionality function in feature calculation for phrases.

4.3 Feature Selection

The last step in engineering our submissions was to select a set of features and a machine learning model from the ones described in Section 3. To do so, we conducted a thorough ablation analysis with all models and multiple feature subsets.

Each feature subset is identified by a set of IDs. Each ID describes a feature or group of features. The identifiers are:

- Word/phrase frequency **(F)**
- N-gram frequencies **(N)**
- Word/phrase length **(L)**
- Number of senses **(S)**
- BERT vector **(V)**

The F identifier represents the (0, 0) configuration described in Section 3.1, while the N identifier represents all others. For example, the subset FNLSV contains all features, while the subset FNS does not contain length or the BERT vector.

The results in Table 3 show the results for the feature configurations that we feel were the most relevant for our engineering process. It can be seen that the best performing variant for both single words and phrases is an SVM trained over all features except n-gram frequencies. Models tend to benefit from the inclusion of word length, number of senses, and especially the BERT vector to the feature set. Interestingly, discarding n-gram frequencies tends to improve the models' performance, especially for single words. This was observed not only in the results of Table 3, but also in many other comparisons we tested, such as FNLSV versus FLSV and FNLS versus FLS.

5 Task Results

We based the creation of the final UTFPR systems on the experiments of the previous section. Our final systems are SVMs trained with word/phrase frequencies, word/phrase length, number of senses, and BERT vector (no n-gram frequencies). Compositionality in phrases was handled through averaging. Frequencies were calculated using a 5-gram language model trained over family movies from SubIMDB. Due to a limitation in time availability, the BERT model was used in its original pre-trained form and not optimized for the task at hand.

Table 4 showcases our shared task performance in comparison to the top 3 and bottom 3 systems with respect to Pearson correlation. Our systems for single words and phrases placed 45th out of 55 and 29th out of 38, respectively. Inspecting the instances that featured most discrepancy between gold labels and predictions, we found that our systems had a tendency of both underestimating the complexity of some of the most complex words and phrases (above 0.7 complexity) and overestimating the complexity of the simplest ones (below 0.2). The conservative nature of their predictions seems to be the main reason why our systems did not place higher.

6 Conclusions

We presented the UTFPR systems submitted to the Lexical Complexity Prediction shared task of SemEval 2021. Although the placing of our systems were not impressive, we do showcase through our preliminary experiments that employing compositionality can potentially improve the predictions for phrases. We also show that including word length, number of senses, and non-optimized BERT vectors to complexity prediction models can noticeably improve their predictions for both words and phrases. In the future, we intend to test the efficacy of adding BERT vectors optimized for the task at hand to the pool of features of our models.

	FN		F		FL		FLS		FLSV	
	Word	Phrase	Word	Phrase	Word	Phrase	Word	Phrase	Word	Phrase
Ridge	0.605	0.585	0.608	0.577	0.602	0.600	0.622	0.603	0.731	0.580
SVM	0.540	0.542	0.597	0.594	0.623	0.525	0.675	0.568	**0.755**	**0.720**
AdaBoost	0.606	0.591	0.603	0.604	0.625	0.591	0.691	0.654	0.710	0.664
GBoost	0.645	0.547	0.597	0.591	0.626	0.532	0.703	0.586	0.732	0.679
Forests	0.579	0.564	0.599	0.469	0.573	0.440	0.684	0.493	0.693	0.673

Table 3: Trial set Pearson correlations for different machine learning models and feature subsets.

System	Words	Phrases
Top 1	0.7886	0.8612
Top 2	0.7882	0.8575
Top 3	0.7790	0.8571
UTFPR	**0.6875**	**0.7601**
Bottom 3	0.1807	0.3197
Bottom 2	0.0971	0.2821
Bottom 1	-0.0272	0.1860

Table 4: Pearson correlation obtained by the UTFPR systems on the shared task compared to the top 3 and bottom 3 systems of each sub-task.

Acknowledgments

We would like to thank the Universidade Tecnológica Federal do Paraná for providing the infrastructure necessary to conduct this research.

References

Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik. 1992. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152.

Marc Brysbaert, Boris New, and Emmanuel Keuleers. 2012. Adding part-of-speech information to the subtlex-us word frequencies. *Behavior research methods*, 44(4):991–997.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*, pages 19–26. ACM; New York.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, 55(1):119–139.

Jerome H Friedman. 2001. Greedy function approximation: a gradient boosting machine. *Annals of statistics*, pages 1189–1232.

Sian Gooding and Ekaterina Kochmar. 2018. Camb at cwi shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194.

Nathan Hartmann and Leandro Borges Dos Santos. 2018. Nilc at cwi 2018: Exploring feature engineering and feature learning. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 335–340.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Tin Kam Ho. 1995. Random decision forests. In *Proceedings of 3rd international conference on document analysis and recognition*, volume 1, pages 278–282. IEEE.

Michal Konkol. 2016. Uwb at semeval-2016 task 11: Exploring features for complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1038–1041.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Shervin Malmasi, Mark Dras, and Marcos Zampieri. 2016. Ltg at semeval-2016 task 11: Complex word identification with classifier ensembles. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 996–1000.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database.

International journal of lexicography, 3(4):235–244.

Gustavo Paetzold and Lucia Specia. 2016a. Anita: An intelligent text adaptation tool. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 79–83.

Gustavo Paetzold and Lucia Specia. 2016b. Collecting and exploring everyday language for predicting psycholinguistic properties of words. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1669–1679.

Gustavo Paetzold and Lucia Specia. 2016c. Semeval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569.

Gustavo Paetzold and Lucia Specia. 2016d. Sv000gg at semeval-2016 task 11: Heavy gauge complex word identification with system voting. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 969–974.

Gustavo Paetzold and Lucia Specia. 2016e. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 1.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.

Matthew Shardlow, Richard Evans, Gustavo Paetzold, and Marcos Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2021)*.

Andrey Nikolayevich Tikhonov. 1943. On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR*, volume 39, pages 195–198.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.